Data Feminism for AI

Lauren Klein∗
lauren.klein@emory.edu
Emory University
Atlanta, GA, USA

Catherine D’Ignazio∗
dignazio@mit.edu
MIT
Cambridge, MA, USA

ABSTRACT
This paper presents a set of intersectional feminist principles for conducting equitable, ethical, and sustainable AI research. In Data Feminism (2020), we offered seven principles for examining and challenging unequal power in data science. Here, we present a rationale for why feminism remains deeply relevant for AI research, rearticulate the original principles of data feminism with respect to AI, and introduce two potential new principles related to environmental impact and consent. Together, these principles help to 1) account for the unequal, undemocratic, extractive, and exclusionary forces at work in AI research, development, and deployment; 2) identify and mitigate predictable harms in advance of unsafe, discriminatory, or otherwise oppressive systems being released into the world; and 3) inspire creative, joyful, and collective ways to work towards a more equitable, sustainable world in which all of us can thrive.

CCS CONCEPTS
• Computing methodologies → Artificial intelligence; • Human-centered computing → Collaborative and social computing theory, concepts and paradigms; • Applied computing → Arts and humanities.

KEYWORDS
feminism, data feminism, data justice, ai ethics, responsible ai

ACM Reference Format:

1 INTRODUCTION
Data Feminism [51] was published in March 2020, in the first week of what would become a world-altering pandemic, and in the wake of over a decade of increasing awareness of the power of data when collected, analyzed, and deployed. Our motivation for writing the book was the overabundance of evidence of the power of data, and of how that power was being wielded unequally. More specifically, it was being wielded by corporations, governments, and other well-resourced institutions to enhance their own power and profit, with significant personal, political, and financial costs for everyone else. There was already a rich scholarly conversation about how data was being used to amplify existing structural inequalities [13, 18, 22, 55, 117, 152]. The contribution of Data Feminism was to show how feminism, because of its analytic focus on the root causes of structural inequalities, could help challenge and rebalance that power. The seven principles of data feminism—examine power, challenge power, rethink binaries and hierarchies, elevate emotion and embodiment, consider context, embrace pluralism, and make labor visible—were intended to operationalize what we saw as the most relevant tenets of feminist thinking for data science. Our goal was to provide a clear set of guidelines and examples for people working with data, who wanted to work with data, or who wanted to refuse to work with data on political or personal grounds. We wanted to show how feminism was not only relevant but essential to data science and, to model how data scientists, computer scientists, digital humanists, policymakers, urban planners, journalists, educators, students, and others could put feminism into practice in their work.

Since the publication of the book, the principles of Data Feminism have been taken up across academia [23, 39, 73, 94, 96, 148, 156] and in the public sector [7, 50, 64, 82, 89, 110]. But as the conversation has shifted from data science to AI, we see a need to revisit these principles. This paper presents a rationale for why feminism remains deeply relevant for AI research, rearticulates the original principles of data feminism with respect to AI, and introduces the possibility of two additional principles, related to environmental impact and consent, in order to help to 1) account for the unequal, undemocratic, extractive, and exclusionary forces at work in AI research, development, and deployment; 2) identify and mitigate predictable harms in advance of unsafe, discriminatory, or otherwise oppressive systems being released into the world; and 3) inspire creative, joyful, and collective ways to work towards a more equitable, sustainable world in which all of us can thrive.

2 BACKGROUND
2.1 What is Feminism?
Feminism has a long, varied, and often contested history. While it exceeds the scope of this paper to summarize the entirety of this history, it is important to clarify the definition of feminism we mobilize here. At its most basic level, feminism entails a belief in the equality of all genders. This includes women and men, as well as Two Spirit, genderqueer, travesti, nonbinary people, and more. But until gender equality is realized in the world, our feminism also requires organized activity to make this goal of equality a reality. A third aspect of our feminism derives from its intellectual heritage, and a crucial part of this heritage is intersectional feminism, which comes to us from the work of women of color feminists, and Black feminists in the United States in particular. The contributions of
intersectional feminism are twofold: first, to bring additional facets of social difference to the conversation about gender inequality, including but not limited to racial and economic inequality; and second to insist that we concern ourselves with structural power: the reasons why people experience privilege on the one hand, or oppression on the other. Intersectional feminism offers models (in the conceptual sense, not the machine learning sense) that explain the causal mechanisms of complex systems of power and guide action to transform them towards justice. These include the Combahee River Collective’s observation about “interlocking systems of oppression” [33], Patricia Hill Collins’s formulation of the “matrix of domination” [76], and Kimberlé Crenshaw’s term “intersectionality” [36]. Note that these intersectional feminist models of power are not the only models of structural power that exist; others have theorized the workings of structural inequality from the perspective of capitalism, colonialism, and so on. We are drawn to intersectional feminist theories of power because of how they bridge personal experience and structural frameworks, and because they are explicit about their goal: understanding present imbalances of power in the world so that they can be challenged, rebalanced, and changed.

2.2 Feminism Today

In the years since Data Feminism was published, there have been several significant alterations to the social and political fabric of the United States and the world, many of which bring feminist considerations to the fore. Most directly, in the US, where the authors of this paper are located, we have experienced the overturning of the constitutional right to abortion, which has set off a cavalcade of increasing incursions into the autonomy and privacy of those with child-bearing bodies. These long-standing feminist concerns now play out in digital (if not exclusively AI-driven) spaces, as personal data has become a key legal weapon wielded against people seeking abortions [135]. Cases such as these also underscore the close relation between reproductive justice and trans justice, in that attacks on the bodily autonomy of some are attacks on the autonomy of all. It is not a coincidence that restrictions on abortion access have been accompanied by restrictions on gender-affirming care, and other legal efforts to police the bodies of trans people. A key feminist lesson here, born from over a century of exclusionary history, is that these are interconnected struggles, and we cannot defend bodily autonomy with an essentialist definition of “women” [137]. A related lesson comes to the US from Latin America. There, the decades-long “marea verde” (green wave) of feminist activism has been successful in large part due to its expansiveness and economic populism as well as its insistence on linking labor issues to gender issues [126]. As the US rolled back abortion rights, for example, Argentina legalized abortion and it was decriminalized in Colombia and Mexico, some of the most populous countries in the hemisphere. These successes were the result of several decades of protest and activism.

Attacks on women and trans people are not happening in a vacuum. Communications scholars W. Lance Bennett and Marianne Kneuer recently argued that we are witnessing the intensification of “illiberal public spheres” around the world [15]. These are characterized by a set of anti-democratic communications tactics that include the denigration and exclusion of minoritized people, targeted attacks on the press and political institutions, and transgression of norms of civil discourse. Such tactics are enabled and amplified by the social media platforms owned by Big Tech, whose business model elevates attention above all other metrics, such that they profit richly off extremism, threats, spectacle, and lies. The success of this corporate partnership with right-wing extremism is evident in the proliferation of book bans [103], the demonization of DEI and critical race theory [37], and mob tactics to doxx, intimidate, misrepresent, and otherwise silence people that study misinformation, speak out against racial and gender violence, or rule against insurrectionists [53]. At the same time, we have seen the continued work of feminists, including Latin American, Indigenous, and abolitionist feminists in the US and globally, doing the work of imagining alternate worlds [38, 43, 127, 139]. This paper represents our own first attempt to imagine a more equitable, sustainable, feminist world for AI research.

3 RELATED WORK

3.1 Feminism and FAccT

Within the FAccT community, we have already seen examples of how feminism can play a substantive role in guiding AI/ML research. As early as 2021, Leila Marie Hampton [68] introduced Black feminism as a lens through which to understand and critique algorithmic oppression. That same year, Hancox-Li and Kumar [70] introduced the concept of feminist epistemology to the FAccT community, offering suggestions on how to align AI/ML research with feminist ideas about the value of situated knowledge [72] and of multiple ways of knowing more broadly. In subsequent years, work at FAccT has demonstrated how a suite of feminist concepts drawn from Black feminism, feminist STS, and feminist new materialism could be applied to ML research [90] and how intersectionality had thus far been (weakly) operationalized in AI fairness research [91].

Feminism has also informed a range of applied work, including research on the harms of online advertising systems [134] and workplace surveillance [28], as well as participatory processes for ML design [145]. Other papers at FAccT have explored the topic of gender more concretely, including the issue of gender bias in public-facing tools (e.g. autocomplete [97]) and research methods (e.g. NLP [48]), as well as in the field of computer science itself [27]. But this work remains a small minority. In their 2022 meta-analysis of AI ethics research conducted at FAccT, Birhane et al. concluded that “the field would benefit from an increased focus on ethical analysis grounded in concrete use-cases, people’s experiences, and applications as well as from approaches that are sensitive to structural and historical power asymmetries” [17]. The principles offered here respond to that call by providing a feminist framework to structure this necessary work.

3.2 Feminism and AI

Looking more broadly, scholarship on feminism and AI stretches back at least to Alison Adam’s 1998 book, Artificial Knowing [3], which, as Keyes and Creel [85] remind us, used feminist epistemology to challenge the presumed universality of AI research of the time. One of the most widely-cited scholarly papers in AI ethics
today, by Joy Buolamwini and Timnit Gebru, employed an intersectional feminist perspective to analyze corporate computer vision systems [22]. As the hype surrounding AI has increased, additional work explicitly linking feminism and AI has emerged. Communications scholar Sophie Toupin [147] conducted a critical survey of this literature to create a typology of six ways that feminism and AI have been linked, including feminist ML models, design-based approaches, and feminist influences on AI policy, culture, discourse and science. A recent edited volume, Feminist AI [20], assembled twenty-one chapters ranging in focus from the dearth of women in AI research [151] to Afrofeminist digital futures [110] to the intersection of AI and racial capitalism [69]. In addition, global feminist networks like the Feminist Internet Research Network [112] and the <A+> Alliance [59] have emerged to support action on issues of algorithmic bias, labor and the economy, AI-induced gender violence, and more. Civil society organizations have also contributed to the conversation about feminist, decolonial and emancipatory approaches to AI. These are too numerous to be comprehensively listed, but some examples include Coding Rights in Brazil [130], IT For Change in India [131], Data Género in Argentina [64], and Policy in Uganda [121]. Taken together, these show the wide-ranging relevance and utility of feminism for AI research.

4 DATA FEMINISM FOR AI

In the sections that follow, we review the seven principles of data feminism and explain how they can be adapted to AI research. We also discuss the possibility of two additional principles that address new considerations brought about by AI’s increasing scope and impact on both people and the planet.

4.1 Principle 1: Examine Power

Data feminism begins by analyzing how power operates in the world.

The first principle of data feminism is to examine power: “the current configuration of structural privilege and structural oppression in which some groups experience unearned advantages—because systems have been designed by people like them and work for people them—and other groups experience systematic disadvantages—because those same systems were not designed by them or with people like them in mind” [51]. When connecting this understanding of power to data science, we focused on issues of unequal power with respect to minoritized groups—and in particular, on the effects of the under-representation of women and other minoritized groups 1 in the field of data science; 2 as the shapers of research questions; and 3 as the subjects of data-scientific research. The predominance of cisgender men – and the exclusion, even banishment, of women (cis and trans) and nonbinary people, as well as Black, Indigenous, and other people of color, especially when they speak out about AI harms – like Timnit Gebru and Margaret Mitchell – is even more acute in AI research and systems development [149]. While Data Feminism touched on the role of corporate interest in determining the focus of data creation efforts and research agendas, we were not yet required to contend with the near-total “capture” of AI research and deployment by corporations that has since taken place [155]. Given this, it is clear that examining power in AI must also centrally involve examining economic power, and the capitalist systems that facilitate the extraction, aggregation, and consolidation of financial resources.

The capitalism at work in the US, the current epicenter of corporate AI research, has been variously named surveillance capitalism [160], oligarchic capitalism [60], and even neofeudalism [44]. Looking back, we can also see its power emerge in the racial capitalism described by Cedric Robinson and others. This is the idea that racialized exploitation has gone hand in hand with capitalist accumulation—indeed that markets are reliant on the production of social hierarchies whereby some groups may own, accumulate, and thrive, and others are excluded, exploited, and marked for premature death. Racial capitalism is also always gendered capitalism, as work by Angela Davis [42], Silvia Federici [57], Verónica Gago [61] and other Marxist feminists have shown. These are systems in which wage gaps, property laws, gender norms, debt structures, and the lack of reproductive rights conspire to maintain women and genderqueer people as a global economic underclass, with additional repercussions for those who are also Black and brown. In short, capitalism is premised upon the preservation of unequal power: of the enforcement of the racial, gendered, and other social hierarchies which enable the extraction of labor, and therefore value, from the many for the profit of the few [35]. These dynamics are clearly visible in the current landscape of AI, in which research agendas are similarly set by the few: it is no surprise, then, that their goal is to maximize corporate profit and to preserve (or even increase) the social and political power that enables it.

Recent scholarship has also drawn attention to the colonial dynamics of AI research [34, 127, 146]. We see this clearly in the outsourcing of low-paid, traumatizing data-labeling jobs to economically vulnerable people in the Global South which we discuss further in Principle 7 on labor. Here, we emphasize how these capitalist and colonialist dynamics leave so-called “externalities”: effects not accounted for in corporate profit models. These include stark environmental impacts, the erosion of job quality, the increased surveillance of workers, a range of harmful incursions on civil rights, outright discrimination, and even death, as we are witnessing right now in Gaza (and discussed in Principle 6, on context). From a macro perspective, when monetary gain is the primary driver of research, we remain in a world in which issues of war, colonialism, racism, and sexism remain unaddressed, since these “externalities” do not contribute to the corporate bottom line.

The artist Mimi O. nu. o. ha’s Library of Missing Datasets [116] underscores this point. For O. nu. o. ha, missing datasets serve as “cultural and colloquial hints of what is deemed important” and what is not. When corporations operate without restrictions or regulations, they determine which issues (or groups of people) are worth collecting data about, and which issues (or people) can be ignored. Shareholder interest and public interest are rarely aligned. Corporate choices determine what research questions can be asked, what analysis can be undertaken, which models can be trained, and which users will ultimately be served by those models. We have seen this play out very visibly with respect to LLMs, where researchers such as Emily Bender, Timnit Gebru, Angelina McMillan-Major, and Margaret Mitchell were quick to point out how the models’ training data was far from representative, even if its size was substantial [12]. Subsequent research has documented more specific
biases against women [92], Muslim people [2], racial and ethnic
minorities [115, 132], specific social roles [101], and more.

Compounding the issue is how these models continue to be
both framed and used as “foundation” models, implying that they
serve as a stable basis on which trusted research can be built. Until
this misconception is corrected, and likely in perpetuity, we must
insist on preventative evaluation, accompanied by an intersectional
analysis of power, as a basic starting point. This is in line with the
US Office of Science and Technology Policy’s Blueprint for an AI Bill
of Rights [114] and the more recent White House Executive Order
on AI, which calls for “robust, reliable, repeatable, and standardized
evaluations of AI systems, as well as policies, institutions, and, as
appropriate, other mechanisms to test, understand, and mitigate
risks from these systems before they are put to use” (emphasis ours)
[77].

These evaluations are necessary for both generative and predic-
tive AI. In terms of predictive AI, in the housing sector, for example,
we have seen how automated tenant screening “services” rely on
eviction records, arrest records, and credit scores in order to rate
tenants on their predicted ability to pay rent. These records are
low-quality due to the non-standardized ways in which they are
published and scraped, as well as the highly racialized nature of the
housing sector in the US [47, 111, 128, 141]. Relying on them dispro-
portionately impacts people of color, and systematically furthers
rather than mitigates structural oppressions.

Here is also where feminist and anti-capitalist critiques of power
converge, pointing to the social and historical causes for why certain
data sources may be biased against certain populations, as well as
to the economic causes for why oppressive and extractive systems
exist in the first place. Put another way: a feminist approach to
examining power in AI must “ask the other question,” as critical
race theorist Mari Matsuda explains, “look[ing] for both the obvious
and nonobvious relationships of domination,” and allowing us to
“see that no form of subordination ever stands alone” [106]. In the
case of the housing sector, it is both power and profit that drives
the informatic asymmetry between the landlords, who can know
almost everything about their tenants because they have so much
more data and AI on their side, and the tenants, who can know very,
very little about their landlords. Tenants do not have the data or the
tools to explore, for example, their landlord’s history of evicting
people unfairly, or of not maintaining their properties. In short,
they do not have the tools to help them build power. A feminist,
anti-capitalist approach to AI would focus on designing systems
in the service of tenant power – and in the service of all of those
excluded by the current rich-get-richer scheme of corporate AI
research and development.

4.2 Principle 2: Challenge Power

Data feminism commits to challenging power and working towards
justice.

The second principle of data feminism is to challenge the unequal
distributions of power that we encounter in the world. In Data
Feminism, we proposed several methods for challenging power
in datasets and data projects, as well as for using data science to
directly confront corporations and governments. These included
collecting counterdata, analyzing data of all kinds with a justice-
oriented lens, imagining alternative ways of doing data science, and
teaching – laying the groundwork for others to continue this work
through data and statistical literacy efforts. In the context of AI,
many of these approaches hold. We still need to respond to political
demands to collect missing data about underrepresented people
and understudied issues. We also need additional ways of devising
models and interpreting their output in ways that work towards
justice. We need to provide holistic AI education that integrates
social and ethical concerns with a healthy dose of history—and
crucially, taught by historians and other humanities scholars—so
that we stop graduating CS students who are overconfident and
underprepared for the complexities of the real world. We need
sound laws and policies to reign in corporate power. Finally, we need
more creative and participatory and democratic ways to imagine
alternative uses of AI, and the space to reject the use of AI if desired
or required.

Examples of each of these aspects to challenging the power of
AI are already underway. For example, in New Zealand, a group
of Indigenous-led researchers are training an ML-backed speech-
to-text system to assist in revitalizing Te Reo Maori, the language
spoken by the Maori people [125]. Meanwhile, LLM researchers are
working to create explicitly multilingual datasets [81] and models
for low-resource languages [98]. Of course, more data collection is
not always the “solution” to problems of inequality. In many cases,
additional data collection can lead to demonstrable harm. This is
the “paradox of exposure” that we name in Data Feminism, “the
double bind that places those who stand to significantly gain from
being counted in the most danger from that same counting (or
classifying) act” (p 105). So as we celebrate these specific examples,
we must also remind ourselves to ask before beginning any new
project whether a technical intervention is even appropriate, as
well as whether we along with the frontline communities we serve
have together considered the range of possible harms [30, 159].

We also need to continue to envision projects that employ AI
in the service of justice, such as generating alt-text for images
to enhance visual accessibility when it is missing [104]. At
the moment, however, there are far too few examples of justice-oriented
AI work. As has been observed, this may be due to the fact that the
data requirements for LLMs and other generative AI models are so
large that they cannot account for small-data approaches [88]. It
might also be due to the fact that these models must be trained on
data from the past, and the past is rife with structural inequalities
that the models learn [29]. Furthermore, these models intentionally
generate output from the center of any probability distribution,
rather amplify lower-probability possibilities. By contrast, feminists
contend that outliers—in language as in life—tell us far more than
data points at the center [49] [154]. This combination of “features” means
that both predictive and generative AI are status quo machines –
truly excellent at reproducing existing conditions and shaping the
future in the image of the racist, sexist, ableist, transphobic past.

Viewed another way, however, the probabilistic basis of these
models can be harnessed to challenge and rebalance power. As
Wendy Kui Hyong Chun helpfully articulates via the example of
models of climate change, these models “offer us the most probable
future, given past and current actions, not so that we will accept
their predictions as inevitable, but rather so we will use them to help
change the future” (emphasis ours) (26). Following Chun, what if we treated the biased output of LLMs not as any ground truth but as but as motivation for intervention so that those biases are no longer experienced in the world?

It is also possible to use both generative and predictive AI to speculate in the Afrofuturist sense—to assist in envisioning otherworldly [26, 32] or alternate futures. For example, work led by Wonyoung So [141] employs reparative algorithms in order to evaluate which possible interventions might be most effective to close the wealth gap between Black and white families in the United States, an example of how AI can work towards rectifying an unjust status quo.

At the same time, the scope and scale of the corporate capture of AI also requires a commitment to collective visioning and collective action. There is valuable work being done on data trusts as an alternative to standard data repositories, such as the Worker Info Exchange, which pools data from rideshare workers so that they can ask questions about their employers [56]. In the LLM space, we might look to the BLOOM model, an attempt at training a fully-documented large language model in an open, collaborative way [158]. Researchers are also considering how they might form independent research coalitions in order to lobby for access to the data of Big Tech that they need in order to conduct their work [105]. And workers whose livelihoods are being threatened by AI are turning to unions and organizing in order to advocate for the conditions they need in order to thrive [8, 153].

These interventions at the level of civil society must be accompanied by government regulation and policy. The EU AI Act, if implemented with force, promises to become a powerful tool to limit the incursion of Big Tech into personal lives, as does the US Executive Order on AI—for as long as it remains in effect. Some countries, such as Canada [25] and Denmark [129], have gone even further, incorporating explicitly feminist and anti-oppressive policies into their governments. We will, of course, always continue to require resistance in the form of care, community, solidarity, and mutual aid—nothing less than a recognition of the shared humanity that capitalism would have us forget. If we are going to resist the complete capture of AI by capitalist forces, then we must return to these emphatically human and anticapitalist models as our guides.

4.3 Principle 3: Rethink Binaries and Hierarchies

Data feminism requires us to challenge the gender binary, along with other systems of counting and classification that perpetuate oppression.

The third principle of data feminism derives from the false binary that Western culture has constructed between the category of “man” and the category of “woman.” There are more than two genders, of course, and a fundamental commitment of feminist thought is to equality for all genders. Furthermore, binaries are often hiding hierarchies, and the gender binary is a key example. It hides a hierarchy-patriarchy-in which cisgender men are on top, dominating social institutions from corporate boards to government leadership positions; women, trans, and genderqueer people are intentionally kept on the bottom; and there is no space at all for anyone in between. One need only look at the skewed gender balance of the signers of the “AI pause” letter, those invited to testify before the US Congress about AI, or the placement of Larry Summers—in famous for his derogatory comments about women and science [74]—at the head of the reconstituted board of OpenAI, in order to see the obvious issues with inclusion in AI. But the issues with the gender binary and with the hierarchy it maintains run much deeper—into the formation of the field of AI research.

The field of data science—the focus of Data Feminism—has never been known for its exemplary inclusion. But the scale and scope of the conversation around the power and perils of data did lead to an observable change—if not a full recalibration—in inclusivity in the field. In under a decade, Harvey Mudd College, under the leadership of president Maria Klawe, increased the percentage of women computer science majors to over 50%, demonstrating that CS professors and administrators that whine about “the pipeline” have no leg to stand on [87, 157]. But it is not a coincidence that as the broader fields of data science and computer science have grown more inclusive, a subset of researchers from within those fields have begun to pull away in order to reconstitute the field of AI. That this group is WEIRD—Western, educated, industrialized, rich, and (arguably) democratic [138]—and also heavily dominated by white cisgender men, appears to us as evidence of another iteration of a common trend in technical fields: the emergence of new gatekeeping mechanisms in order ensure that those at the top are able to maintain their elite status in the field.

We also see this power move taking place on a conceptual level, in which consumers of AI technologies are being encouraged to believe that its capabilities are magical and mysterious, and thus impossible for any normal person to understand. This, too, has a historical antecedent in the way that software developers became figured as “wizards” and “sorcerers” in a way that made the acquisition of technical expertise seem off-putting to those without access to formal training mechanisms [54, 75]. But it also serves a strategic purpose: it makes it appear to end-users that they should not question or make demands of such tools; to outside researchers that they should not demand any documentation or ability to peek under the hood; and to government regulators that they lack the knowledge to draft policy or legislation to govern AI. We run the risk of repeating this pattern when we accept the language of AI’s “emergent properties” and the hagiography of the field’s supposed founders, as echoed in a 2023 New York Times article that elevated exclusively men, mostly ultra-rich and white [86]. This is why the inclusion implied by the slogan of the Algorithmic Justice League is so important at this present moment: “If you have a face, you have a place in the conversation.” [93].

It is also important to recognize the very real ways the gender binary has been operationalized and even weaponized in AI research. Beyond ample evidence of the gender biases against cisgender women, and trans and nonbinary people, that are entrenched in LLMs, AI researchers appear to have an antiquated understanding of gender itself [48, 84, 136, 142]. This lack of understanding of gender, in turn, leads to the inaccurate (at best) and harmful (at worst) research design and applications. And sex fares no better [5]. There is a near-universal failure to understand that neither sex nor gender are essential, “natural”, or fixed properties, and they certainly cannot be “detected.”
4.4 Principle 4: Elevate Emotion and Embodiment

Data feminism teaches us to value multiple forms of knowledge, including the knowledge that comes from people as living, feeling bodies in the world.

Elevating emotion and embodiment with respect to data is connected to the previous principle, because emotion is often viewed in binary opposition to reason, as a "feminized" element that should be excluded from scientific research because it is subjective and emotion are in opposition to objective, "soft" (feminine stereotype) instead of "hard" (masculine stereotype). With a critique of hidden hierarchies in mind we can also return to the binary between reason and emotion and recognize 1) how it is a false binary, and there is no such thing as purely rational science [83]; and 2) how this binary is hiding a hierarchy, in which emotion, embodiment, lived experience, and other feminized ways of knowing are relegated to lower status forms of knowledge.

The tools to challenge this false binary come from feminism. Feminist philosopher Donna Haraway’s concept of situated knowledges—the idea that knowledge originates at a particular time, in a particular place, and from within a particular set of social and political contexts—helps us recognize how all knowledge, including scientific knowledge, is shaped by the perspectives of the people who produce it [72]. More recently, Black feminist theorist Katherine McKitterick points us to how there are multiple systems of knowledge-making, and to how these must be understood as sites of "collaborative praxis" in and of themselves, and as relational with respect to one another [107]. Crucially, a refined awareness of multiple knowledge systems enhances, rather than detracts from, our collective knowledge.

This feminist objectivity bears relevance to AI research in myriad ways, not the least in how the conversation around the risks and harms related to AI has unfolded over the past year or so. On the one hand, we have seen those in positions of power (who, not coincidentally, themselves represent dominant social groups) sound the alarm on future hypothetical harm—of AI robots developing nuclear weapons and other questions of moral “alignment.” On the other, we have seen AI researchers with equal technical expertise, enhanced by the “empiricism of lived experience” as women (cis and trans), queer people, and/or people of color, calling attention to the harms being perpetuated by AI systems right now [66]. Famously, Dr. Joy Buolamwini, a Black woman computer scientist, undertook an “evocative audit,” and used the example of her own face to demonstrate the poor accuracy of facial recognition software. She then connected that evidence to the greater harms being unleashed as these inaccurate software systems were (and remain) deployed in police departments which are already inextricable from structural racism [21]. We have also seen research documenting the misogyny and racism, as well as instances of rape and pornography, in multimodal datasets such as LAION, resulting in LAION pulling its datasets from circulation [65]. Very crucially, this research was first performed by Black women-led research teams who brought their personal experience with these datasets to their research. Yet it was only when this research was replicated by white researchers at Stanford that the team behind LAION took note. This is unfortunately already a pattern in AI research in which Black women who bring their embodied experience to their research have their concerns dismissed, only to have those concerns validated by people from dominant groups, often years later [12, 149].

It is not just that the full range of human experience should be incorporated into assessments of AI harms. Another binary that must be challenged has to do with our perception that work with data and AI can only serve to increase efficiency, automation, and control. Recent research describing the data practices of feminicide data activists has shown how the production of data can become a tool of intimacy, relationality, care, and memory work [63]. Scholar and activist Helena Sánchez Val has described these works as “affect amplifiers,” translating feminist grief and rage into public action [150]. This form of collective, community-engaged, intentionally emotion-laden work can also remind us of our interdependence, a key idea from disability studies that emphasizes how “we rely on each other, and our actions can have consequences on others” [9]. As we move forward, we must continue to consider how to craft AI systems that focus on “practicing alliance,” create “carewebs and pods,” and are designed for “misfitting.” [46, 62, 120]. Emerging steps towards such systems are happening in human computer interaction with conversations around trauma-informed computing, healing databases, and restorative/transformative data science, and should be expanded to AI research.

4.5 Principle 5: Embrace Pluralism

Data feminism insists that the most complete knowledge comes from synthesizing multiple perspectives, with priority given to local, Indigenous, and experiential ways of knowing.

The principle of embracing pluralism builds from the previous principle, which emphasizes alternate forms of knowing. Here, the emphasis is on how we might represent multiple perspectives in our work. Crucially, “multiple perspectives” does not simply mean multiple opinions. Building on the work of Sandra Harding, and more recently, the Design Justice Network, “multiple perspectives” has become a beacon for the wide range of experiences, social positions, and places in the world from which people produce knowledge. The central premise of embracing pluralism is that we can gain better, more detailed, more accurate, and ultimately more truthful
knowledge if we pool these perspectives together, paying particular attention to the perspectives of those who are most directly impacted by the issue at hand.

Embracing pluralism is of crucial importance in the context of AI research, both because of the narrow demographic composition of AI researchers themselves (see Principle 3, on binaries and hierarchies), and because of the imbalance of power between those currently designing AI systems, and those subject to their decisions (see Principles 1 and 2, on power). Put simply: the field of AI needs to develop more participatory, more responsible, and more humble methods for being in dialogue with impacted communities. Catherine had the opportunity to learn this first-hand in a project involving the co-design of AI tools in collaboration with grassroots feminist data activists across the Americas. Together, they decided to develop an ML classifier for detecting news articles about feminicide, a systematic violation of human rights that involves the gender-related killings of women and girls. The tool was intended to streamline the work of the activists. But the activists pushed back on fully automating the process for two reasons: (1) the impossibility of generalizing any definition of feminicide, and (2) because of how they saw their work as a form of memory justice: a way to honor the lives of the women killed so that they would not be forgotten. The result was a tool that has seen impressive uptake by human rights monitoring groups around the globe. This stands in contrast to the numerous ML/AI tools created without consultation, which lack both users and use.

The prospect of embracing pluralism becomes harder in terms of both architecture and access when scaling up to larger models. LLMs must be pre-trained by those with access to both data and compute, introducing barriers to early-stage participation that had previously not existed. While the push for more open-source models (e.g. BLOOM, OLMo) and more transparent models (e.g. the Stanford Transparency Index) are certainly welcome, the technical requirements of training such models make it far more difficult to involve impacted individuals and small groups in their creation. One recent example that suggests a path forward comes from a project led by Maria Antoniak [10] which surveyed both pregnant people and healthcare providers about their ideas and fears about the possibility for an AI-assisted chatbot that might help to navigate the experience of pregnancy. Like the feminicide classifier, Antoniak et al. very crucially did not build the chatbot first and then ask for feedback; instead, they created a study that included people with multiple forms of expertise and that enabled them to share that knowledge. The output of the project was a set of values, distilled from the participants themselves, that might guide future work.

There is another level at which we might conceive of pluralism, which is at the scale of government and those democratic bodies which have been designed to serve the public interest. We see this approach to embracing pluralism in the ideas of public-interest AI [19] and digital public infrastructure [161]. Like physical infrastructure – roads, parks, schools – these systems would be conceived and designed through public processes, guided by the public interest, and come with transparency and governance requirements. While we have seen important lawsuits and other regulatory processes directed at corporations and the technologies they develop, we might also envision citizen-led design processes in which ideas are sourced from communities and projects are funded (and maintained) with taxpayer dollars.

These examples of participatory design notwithstanding, it is important to acknowledge that “participation is not a design fix for machine learning,” as Mona Sloane et al. have explained [140]. Participation can be tokenizing, extractive, under-resourced by researchers inexperienced with the investment required, or incorporated too late in the design process to influence its outcomes, among other issues that the authors raise. We must continue to push back against what Lorraine Daston might call the “epistemic virtues” of AI/ML research, which echo those of the larger field of computer science: its emphasis on novelty, generalizability, and efficiency, which not only over-determine the kinds of research that are support but are also implicitly believed to be best [40]. While much attention has been devoted to the capture of technical research by corporations, it is also true that there exists a capture of academic research by the field of computer science. Those of us in academia have observed the trends in the composition of college majors, and resultant faculty hiring, leading to outsized resources being invested in computer science and the forms of research it supports to the detriment of all other fields.

Computer science continues to push out women and people of color at a scale not seen in other sciences. It systematically fails to teach the value of context-sensitive, culturally appropriate technology development, and engages in masculinist fantasies of foundational models ("one model to rule them all"). The singular hubris of this field seems to know no bounds, even as evidence of harm surfaces over and over again. We have also observed how fields such as HCI and social computing draw from expertise in, for example, anthropology, sociology, and literary studies, even as CS departments refuse to hire scholars with PhDs in those fields to train their students. If we are sincere in our belief that the best, most accurate, most truthful knowledge comes from the pooling of multiple perspectives, then we must restructure our own research processes and funding models in order to account for this fact.

### 4.6 Principle 6: Consider Context

Data feminism asserts that data are not neutral or objective. They are the products of unequal social relations, and this context is essential for conducting accurate, ethical analysis.

The sixth principle, to consider context, is at once universally applicable to AI research and yet (nearly) universally ignored. This principle applies most directly to the issue of training data, which we began to discuss with respect to language models in the principle on power. Related work has shown how additional (human) decisions made during the training and filtering process introduce additional biases into these datasets, such as for text written by those of higher socioeconomic status [67], and those who occupy specific social and professional roles [101]. Related research has explored text-to-image models, with findings related to the reproduction of sexist, racist, and colonial biases [113, 124]. What we learn from feminism is how understanding the contexts in which these datasets are created, and more broadly, a recognition of how all data is shaped by unequal social relations, is essential for identifying any downstream biases or potential harms. It is also necessary
for ensuring that research questions are properly framed, that evidence is properly analyzed, and that any claims that might be made on the basis of that analysis are properly scoped.

There exist several valuable recommendations for how to restore context to ML models more broadly, such as Eun Seo Jo and Timnit Gebru’s suggestion to adopt practices similar to those employed by archivists when curating and documenting datasets [80], and Mitchell et al.’s proposal of “model cards for model reporting” when releasing a model for general use [108]. We would be well-served by considering how these might be adapted to the current AI/ML landscape, particularly given that neural-network-based architectures of generative AI models make it difficult for the end user (or anyone) to trace any specific model output back to the sources that contributed to it. But to do this work, we must challenge the present stratification of labor in the data science pipeline – the erroneous idea that curating, labeling, and documenting data is somehow unskilled labor and that the analysis and modeling part of the pipeline is where the “science” is at [58]. As archivists and others in the humanities know well, the work of creating a dataset that accurately represents the research question at hand is long, painstaking, and premised on deep expertise. Yet the dominant mantra of Big Tech remains “move fast and break things.” This fundamental mismatch of value folds back into the capitalist critique that began this paper; so long as the production and curation of data is not seen as valuable, then we will not be able to sufficiently support this type of human expertise. Moreover, the AI systems that we produce with naive and ill-considered data will just plain get things wrong. These are the “data cascades” that Sambivasa et al. identify as amplifying into major quality problems downstream [133].

A final consideration with respect to context has to do with historical context—and in particular, of the historical context of AI research itself. This history is an uncomfortable one, as it points to the longstanding complicity of the field of computer science with US military research. This dates back to the DARPA-funded creation of ARPA.NET, the precursor to the internet [4]. It carries through to the construction of OntoNotes, the hand-labeled language model underlying many common NLP libraries, which was another DARPA-funded project [123]; and to specific approaches to language modeling, such as topic modeling, which came about through a government desire to monitor global newswire messages at scale [16]. More recently, news reports have drawn our attention to the role that Amazon has played in cloud storage for the US Immigration and Customs Enforcement agency (ICE) [71]; and how Google developed technology to make drone strikes more accurate, which they then sold to the US government [118]. These developments are no longer theoretical, as evidence emerges that they are being put to use by Israel against Palestine, and with horrific human costs [41]. As advocacy efforts such as #NoTechForApartheid [11] and publications such as Logic(s) Magazine [1] powerfully remind us, these inhumane, destructive, and—in the context of Palestine, genocidal—contexts are those in which too many technical innovations are put to use. We, as AI researchers, can no longer claim ignorance about these contexts as among the uses for our work.

4.7 Principle 7: Make Labor Visible

The work of data science, like all work in the world, is the work of many hands. Data feminism makes this labor visible so that it can be recognized and valued.

This final principle highlights the many people whose labor enables work with data. In Data Feminism, we observed how the work of data science replicated professional hierarchies, with credentialed data scientists at the top, and those perceived to occupy less technical roles—such as data annotation and content moderation—on the bottom. We also observed how this professional hierarchy could be mapped onto gendered, raced, and ultimately colonial hierarchies, with those in the Global North occupying the high-status and high-compensation roles, and those in the Global South occupying those at the bottom.

Because the current configuration of AI research is premised upon the consolidation of significant resources—technical and economic as well as human—this colonial structure has become solidified as the fundamental framework on which AI depends. One need only look at the investigative reporting that followed the initial release of ChatGPT, which showed that this “artificial” intelligence depended on very human workers in Kenya screening potentially offensive responses in real-time [119]. The location of these workers is not a coincidence. Scholars such as Julian Posada have asserted that companies in the Global North exploit political instability and capitalize on catastrophe to enrich themselves, a familiar and longstanding historical pattern [122]. This joins a long line of research (and evidence in the world) that documents how capitalism is fundamentally dependent upon resource extraction—and on paying as little as possible for those resources, including human labor, in order to maximize profit [35, 127].

The workforce of the Global North is not exempt from the incursion of AI, of course. In the US, we have seen lawsuits by Getty Photography and The New York Times brought against companies peddling generative AI technologies. In addition, the summer and fall of 2023 saw major strikes by the Writers Guild of America [8] and the Screen Actors Guild [153]. These culminated in necessary protections against the use of AI to revise scripts and in the creation of likenesses of human actors. We also celebrate the larger trend towards collective action in the tech sector, and in white collar jobs more broadly, as a counterforce to the relentless individualism championed by capitalism. Graduate students at elite engineering universities in the US are beginning to unionize, joining long-standing unions at public institutions including the University of California system, the Cal State system, and the City University of New York, among others. We must continue to ensure that these efforts at building solidarity cut across lines otherwise drawn by technical expertise. After all, the history of colonialism tells us that those at the lower end of labor hierarchies will be the ones most impacted by any move to increase profit or workplace efficiency. Without solidarity across class, race, gender, and work sector, capitalist power will only continue to accrue.
5 DISCUSSION: FUTURE PRINCIPLES FOR FEMINIST AI

The previous pages document our current thinking about how the seven principles defined in Data Feminism can be applied to AI research, but two topics require additional attention: the environmental impact of AI research and deployment, and issues surrounding consent. Here we summarize our present thoughts on each.

5.1 Data Feminism and the Environment

In many ways, questions about the environmental impact of AI follow from how its development and deployment reinforce historical patterns of capitalism and colonialism. Resource extraction, after all, is as much about natural resources as it is about human labor. It has long been observed that the environmental impacts of this resource extraction are experienced unequally, with people in the Global South experiencing the deleterious effects of climate change in far greater measure than those in the Global North, even as they contribute far less to global emissions. Google, for example, used 5.6 billion gallons of water in 2023, up 20% from the prior year [78]. An average Meta data center consumes as much electricity as 150,000 average homes [109]. As current research into the energy and water requirements of LLMs has shown [99]. AI seems positioned to further exacerbate these effects. Again, these systems seem positioned to benefit elite users in the Global North, even as they exact their cost on those in the Global South. This is an environmental issue, but it is also a feminist issue, as these effects are not only experienced unequally in terms of geography, but also in terms of gender. A feminist principle for AI about the environment might draw from the several decades of ecofeminist scholarship which has worked to establish the connection between environmental harms and other forms of structural oppression. It might also look to work by Indigenous feminists in Latin America, who view the “cuerpo-territorio” (body-land) as an interconnected system, and by North American Indigenous feminists who similarly link body and land sovereignty while working towards the end of structural violence against both [24, 45, 100, 139].

5.2 Data Feminism and Consent

Consent is also a longstanding feminist concern due to the high rates of rape and sexual violence faced by women, trans and non-binary people around the world living under cisgender patriarchy. Most Western laws that address rape and sexual violence have their basis in some form of consent, and there are various feminist formulations of what that might mean, such as the popular 2016 FRIES model from Planned Parenthood where consent is: Freely given, Reversible, Informed, Enthusiastic, and Specific. With that said, there are numerous feminist critiques of consent as being too individually focused, too simplistic, and too binary (e.g. reinforcing heteronormative, gendered stereotypes of aggressive men and gatekeeping women and ignoring queer relations entirely) [6, 52, 102]. Because of the violence and harms already being propagated by AI systems, and because of the likelihood that these harms will continue to increase, we still think it may be useful to formulate a feminist principle for AI about consent in expanded forms, i.e. queer, collective and/or interdependent consent.

AI is currently facilitating the mass, non-consensual exploitation of pornographic images of women in the form of deep fakes. As Danielle Citron teaches us, this is consistent with a number of “intelligent” technologies such as networked cars, mobile phone apps, and more which have been exploited by abusive perpetrators to control and dominate their partners[31]. Then, there are the broader issues of consent that relate to the data sources used to train LLMs and generative AI systems, as discussed in Principle 6. Are online data—including social media posts, family photos, original artwork, journalistic reporting and personal blogs—fair game for inclusion in massive data sets without the creators’ knowledge? As we await the development of informed guidelines for fair use, we can be certain that something other than the current system—in which Big Tech steals people’s work, exploits it, makes money, and facilitates structural violence along the way—is required. There has already been significant work around consent and technology, including the Consentful Tech Project by Allied Media Projects, and work on what consent means in human computer interaction [79, 95, 144] which we can use to build more relational, inclusive, and liberating systems—and to reject them if they are not.

6 CONCLUSION

While the forces of racial, gendered capitalism that are currently shaping AI research are powerful, they are also predictable. They operate in ways we have observed and experienced for centuries. We offer these thoughts on feminist principles for AI research, along with our hope, because we know what will happen if we do not change course. The predictability of late-stage capitalism, in some ways, gives us an easy goal: if the status quo is not what we want, then we must follow Ruha Benjamin’s call to “craft the worlds you cannot live without, just as you dismantle the ones you cannot live within” [14].

ACKNOWLEDGMENTS

We would like to thank Nikki Stevens and Isadora Cruxên for their feedback on early drafts of this paper. This work has been partially supported by a grant from the Mellon Foundation (G-2211-14240).

REFERENCES

We are a two-person writing team that brings domain expertise in the humanities (including historical and literary scholarship), digital humanities (including NLP/ML), urban planning, software development, data science, and data visualization. As two cisgender women, we share an interest in and commitment to gender equality and social justice. As researchers for citational justice, and we have specifically sought to cite the work of scholars from marginalized backgrounds, especially BIWOC and queer people of color, as well as forms of knowledge production not recognized by the academy, including activism, journalism, art, design and creative communication projects.

A.2 Research Positional Statement

We are a two-person writing team that brings domain expertise in the humanities (including historical and literary scholarship), digital humanities (including NLP/ML), urban planning, software development, data science, and data visualization. As two cisgender women, we share an interest in and commitment to gender equality and social justice. As researchers for citational justice, and we have specifically sought to cite the work of scholars from marginalized backgrounds, especially BIWOC and queer people of color, as well as forms of knowledge production not recognized by the academy, including activism, journalism, art, design and creative communication projects.

A.3 Adverse Impact Statement

With this work, we seek to spark conversations across disciplines about the social and political inequalities being exacerbated by AI. Rather than “calling out” specific people or institutions, we hope this critique serves as a “call in” other scholars to join together and work collectively towards building the AI systems that we all deserve.

submitted 22 January 2024; accepted 30 March 2024